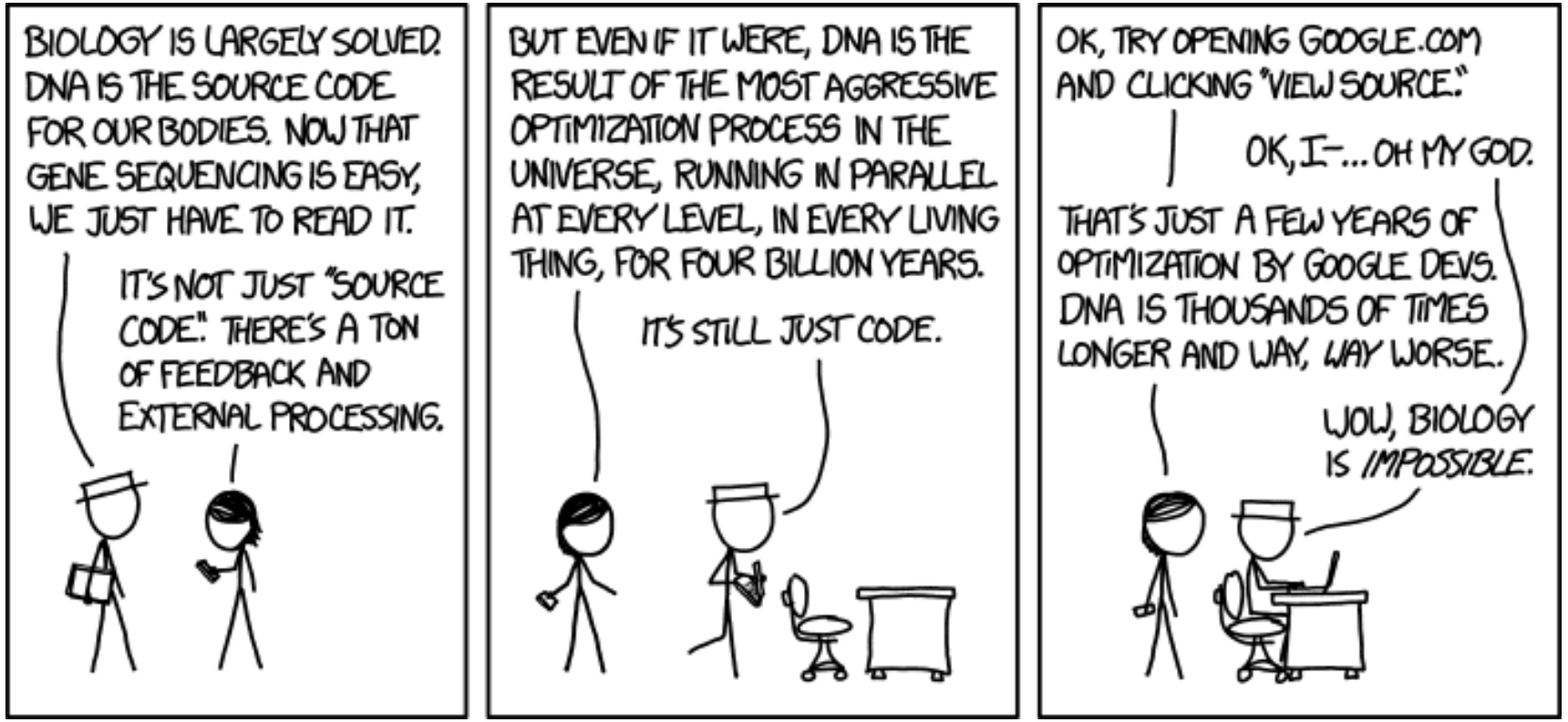
AMIDD Lecture 2: Biological Sequence Analysis



DNA by Randall Munroe, https://xkcd.com/1605/

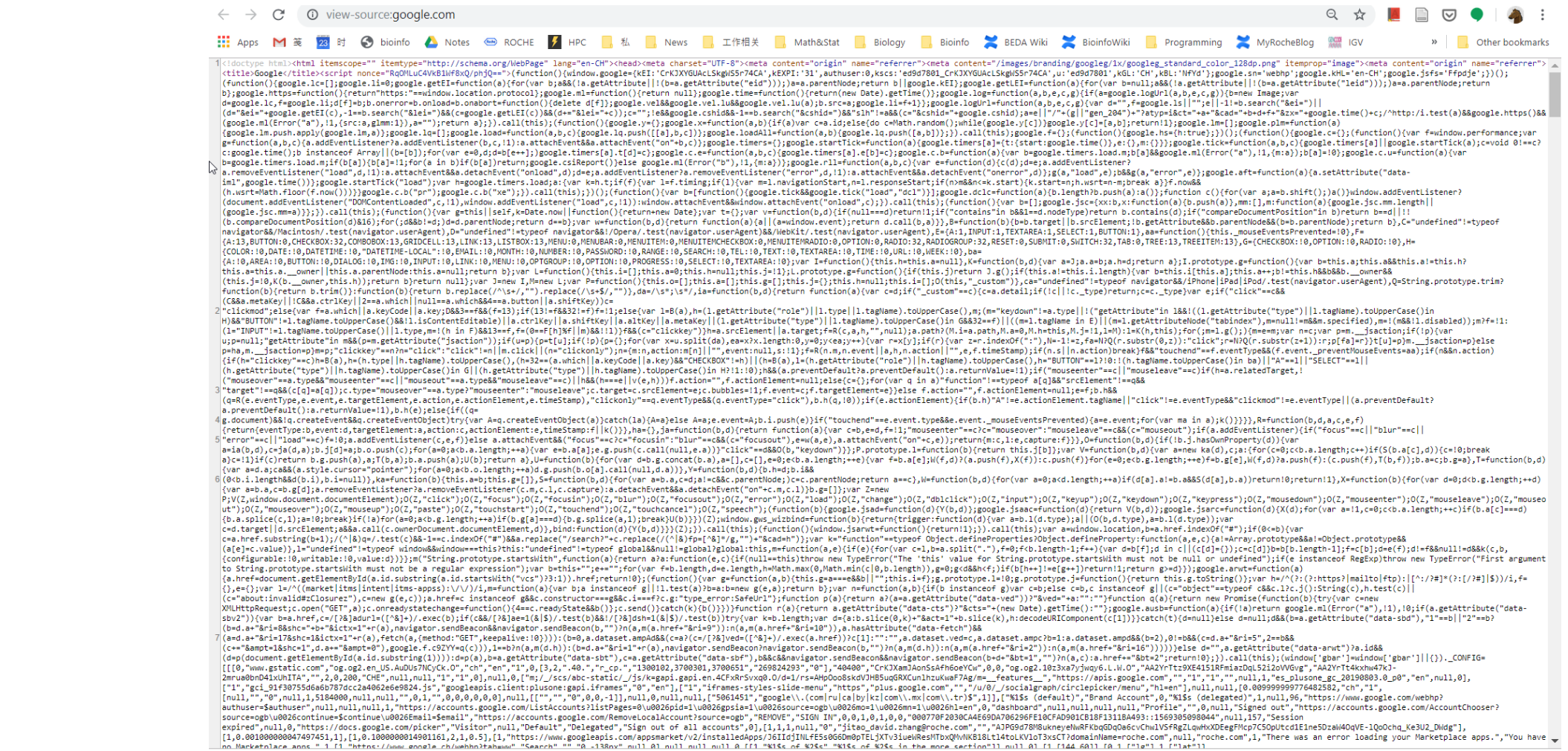
Dr. Jitao David Zhang, Computational Biologist

[1] Pharmaceutical Sciences, Pharma Research and Early Development, Roche Innovation Center Basel, F. Hoffmann-La Roche

[2] Department of Mathematics and Informatics, University of Basel

Part of the source code of *Google.com*

As of 24.09.2019



Contact the author

- The central dogma of molecular biology lies in the core of modern drug discovery
- Biological sequence analysis is used to
  - Understand encoding of biological information
  - Compare between genes and between spcies
  - Develop new drugs
- We will highlight Dynamic Programming and the Markov Chain as examples

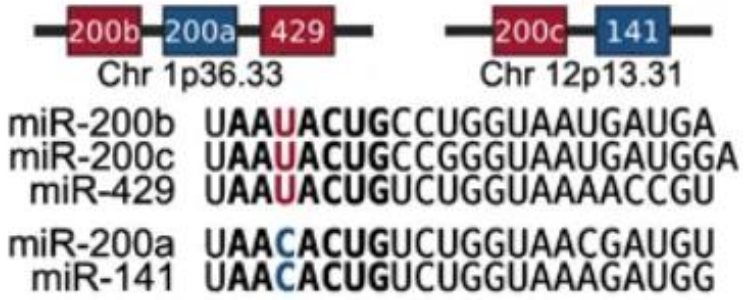
Questions on the video on Herceptin by Susan Desmond-Hellmann

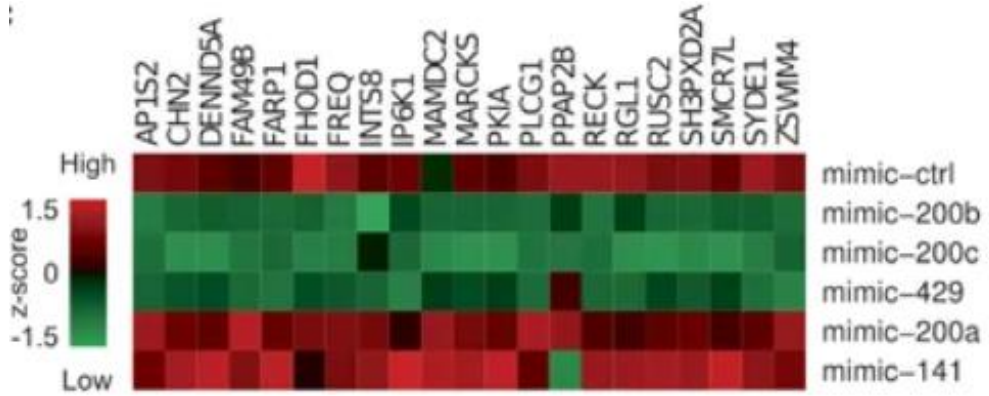
1. What is the **indication** of *Herceptin*? What is its generic (USAN, or United States Adopted Name) name?
2. What is the **gene target** of Herceptin?
3. In which year was the **target** of Herceptin described? When was Herceptin **approved**?
4. What was the **improvement** of Herceptin compared with earlier antibodies?
5. Why does a **biomarker** matter besides developing drugs?
6. In the clinical trial of *Herceptin* for **metastatic breast cancer**, how much improvement in the **median survival** did Herceptin achieve? And how much improvement is in the **adjuvant setting** (Herceptin applied directly after operation)?

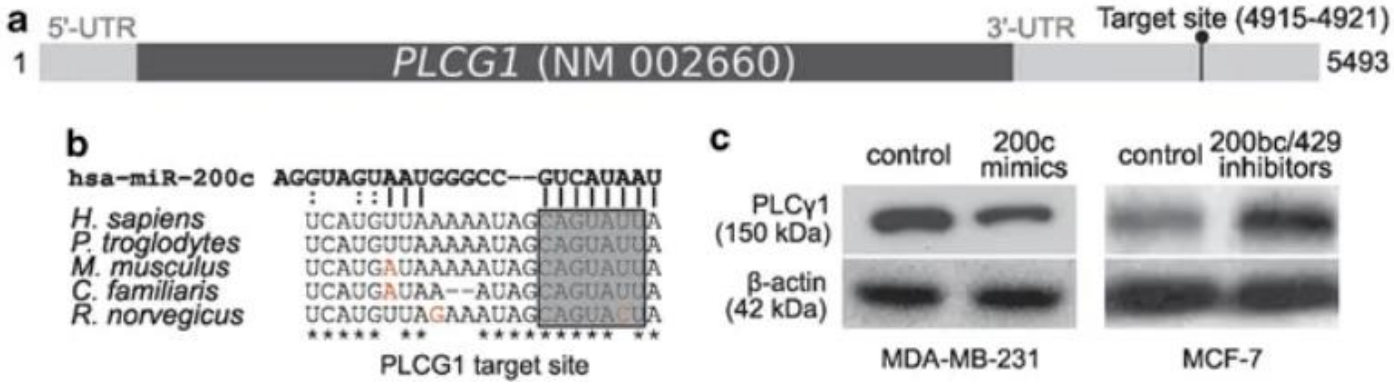
Questions for further thinking

- Susan Desmond-Hellmann summarizes great drug development in four key concepts: (1) Having a deep understanding of the basic science and the characteristics of the drug. (2) Target the right patients. (3) Set a high bar in the clinic. (4) Work effectively with key regulatory decision markers. Where do you think mathematics and informatics play a crucial role?
- She emphasized the importance of collaboration. What skillsets do we need for that?
- How do you like her presentation? Anything that you can learn from her about presentation and story telling?

A single-nucleotide difference can lead to huge difference







Uhlmann *et al.*, Oncogene, 2010

A single nucleotide change leads to distinct effects of miR-200bc and -200a in breast cancer cells

Questions about Bollag *et al.,* Nature 2010

1. What is the **indication** of *PLX4032*?
2. What is the **gene target** of *PLX4032*?
3. The maglinancy depends on which biological **pathway***?*
4. What is the **Mechanism of Action** of *PLX4032?*
5. What went wrong in the first **Phase I clinical trial**? And how was it solved?
6. What was the **dosing regimen** in the final Phase I clinical trial, and what is the **response rate**?

Questions for further thinking

- Susan Desmond-Hellmann summarizes great drug development in four key concepts: (1) Having a deep understanding of the basic science and the characteristics of the drug. (2) Target the right patients. (3) Set a high bar in the clinic. (4) Work effectively with key regulatory decision markers. What parts of this abstract reflect these points?
- Susan Desmond-Hellmann emphasized the importance of collaboration. Is that true when you consider this abstract?
- How do you like the abstract? Anything that you can learn from it about writing?

A single-amino-acid difference in BRAF gene may mean longer survival of melanoma patients given the correct treatment

McArthur, Grant A., Paul B. Chapman, Caroline Robert, James Larkin, John B. Haanen, Reinhard Dummer, Antoni Ribas, *et al.*

Safety and Efficacy of Vemurafenib in BRAFV600E and BRAFV600K Mutation-Positive Melanoma (BRIM-3): Extended Follow-up of a Phase 3, Randomised, Open-Label Study

The Lancet Oncology 15, Nr. 3 (1. März 2014): 323–32. https://doi.org/10.1016/S1470-2045(14)70012-9.

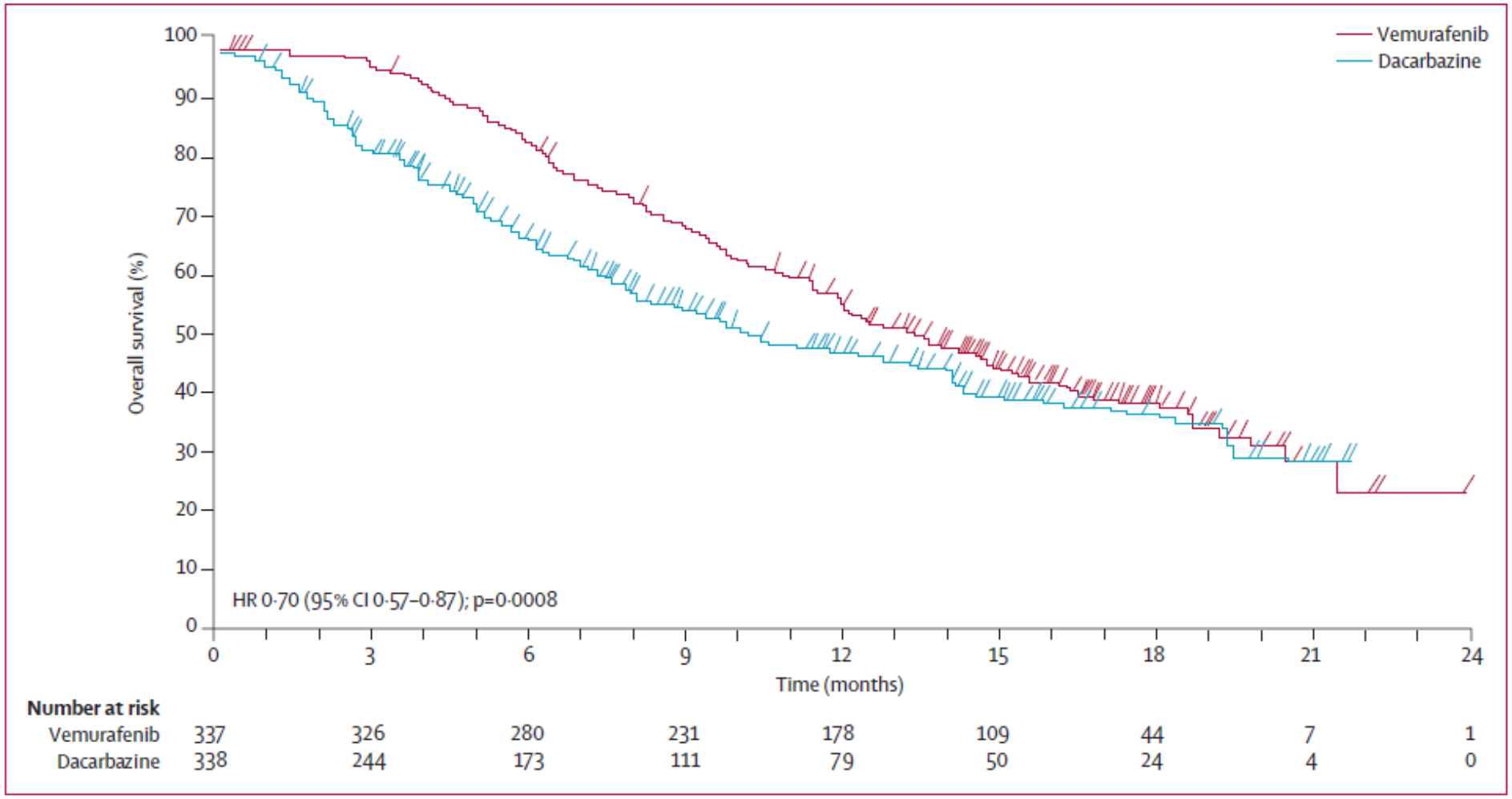


Figure 2: Overall survival (randomised population; censored at crossover) for patients randomly assigned to vemurafenib or to dacarbazine (cutoff Feb 1, 2012)

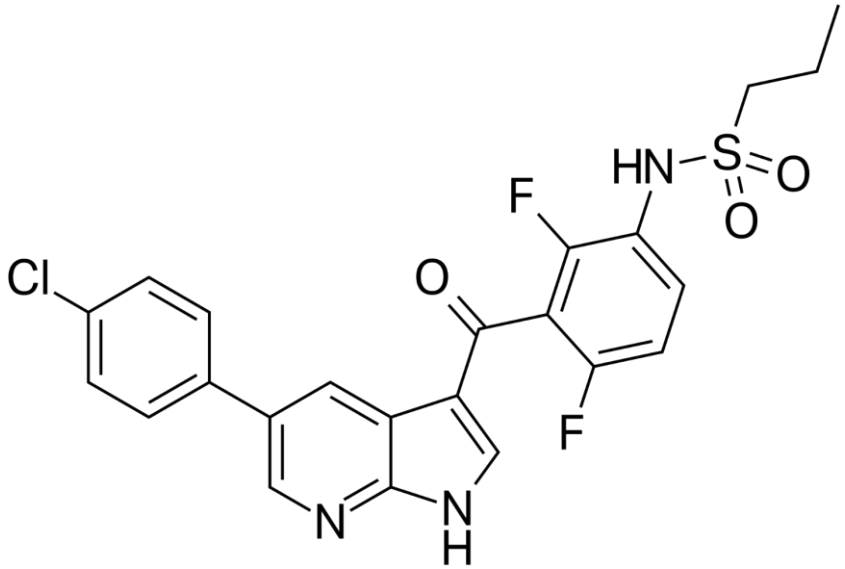
Vemurafenib (Zelboraf, PLX4032)
V600E mutated BRAF inhibition

- V600E: Valine (V) on the amino-acid position 600 is substituted by glutamic acid (E).

```
                560         570         580         590         600
IKLIDIARQT AQGMDYLHAK SIIHRDLKSN NIFLHEDLTV KIGDFGLATV
                610         620         630         640         650
```

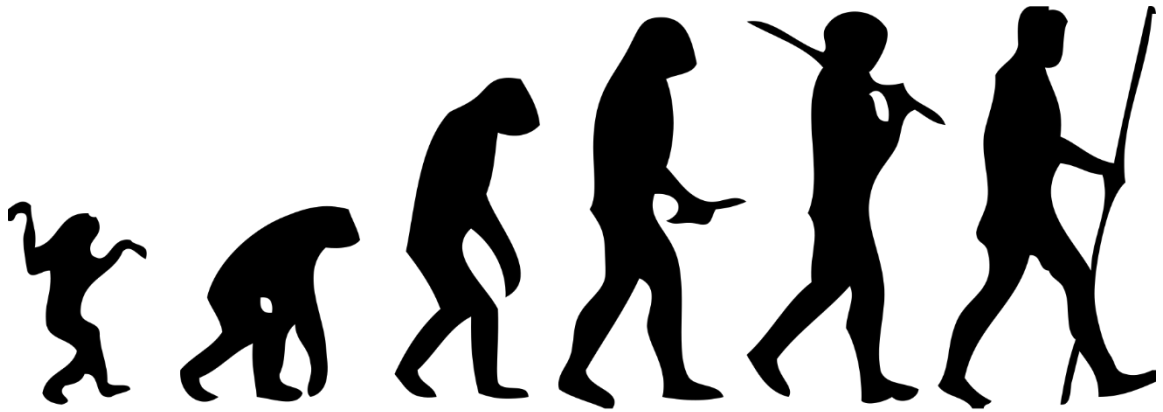
Fragment of BRAF protein. Source: UniProtKB, P15056 (BRAF_HUMAN)

- View the 3D structure of the molecule at PDB ligand database
- View the X-ray structure of BRAF in complex with PLX4032 on PDB: accession number 3OG7.
- Find more information about the discovery and clinical efficacy of vemurafenib in the handout.

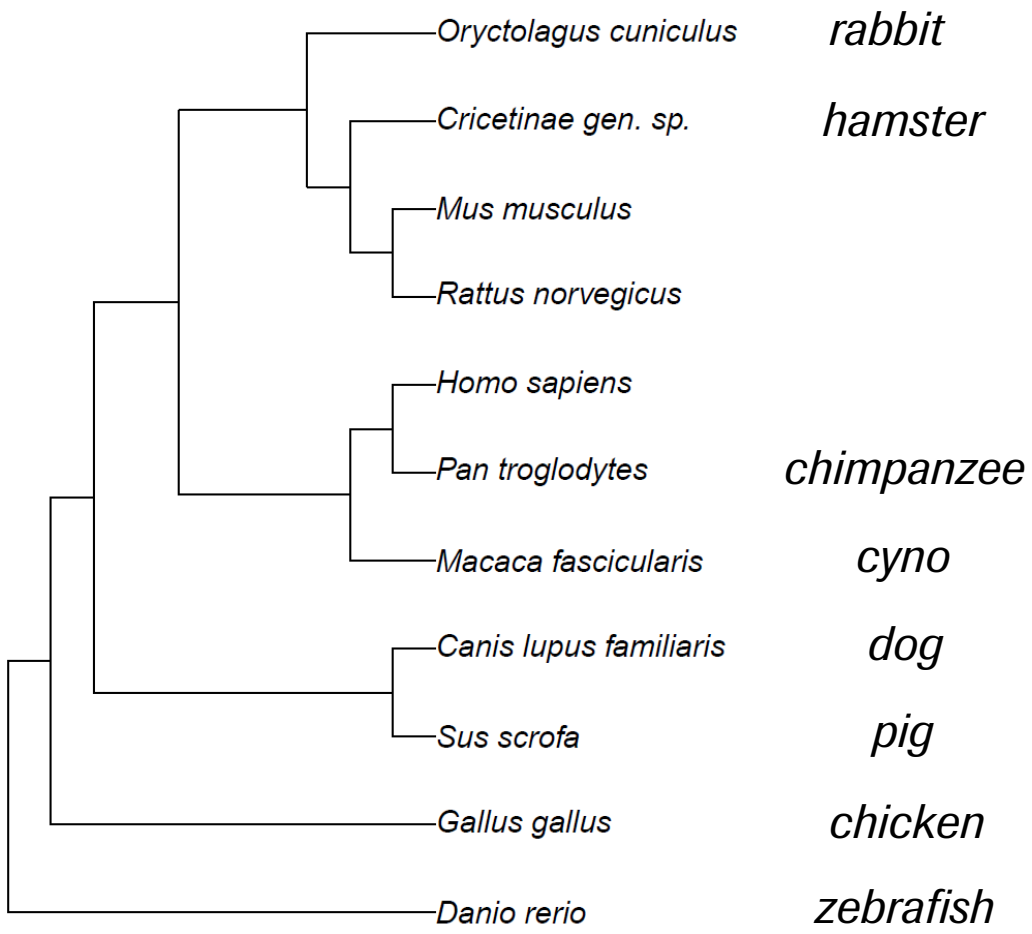


Source: https://commons.wikimedia.org/wiki/File:Vemurafenib_structure.svg

Evolution



Phylogeny of commonly used species for animal studies



Tree structure retrieved from https://itol.embl.de/ (iTOL, Interactive Tree of Life), visualized with the *FigTree* software by Andrew Rambaut

DNA

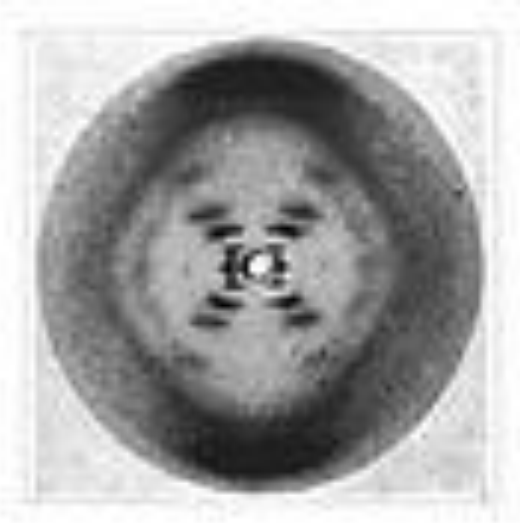
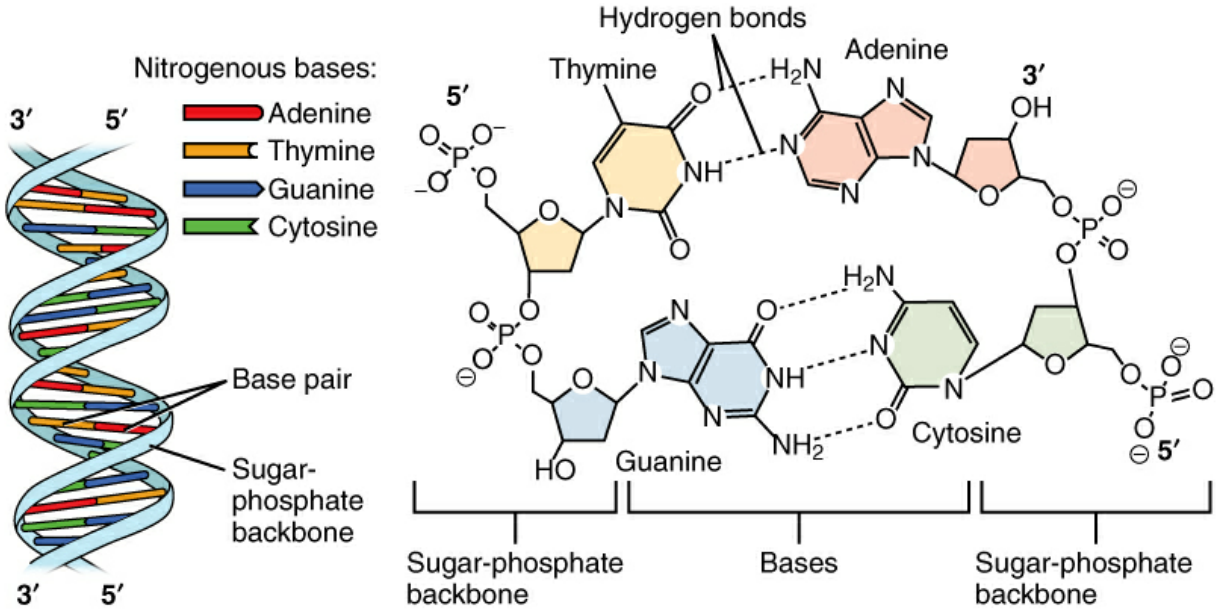
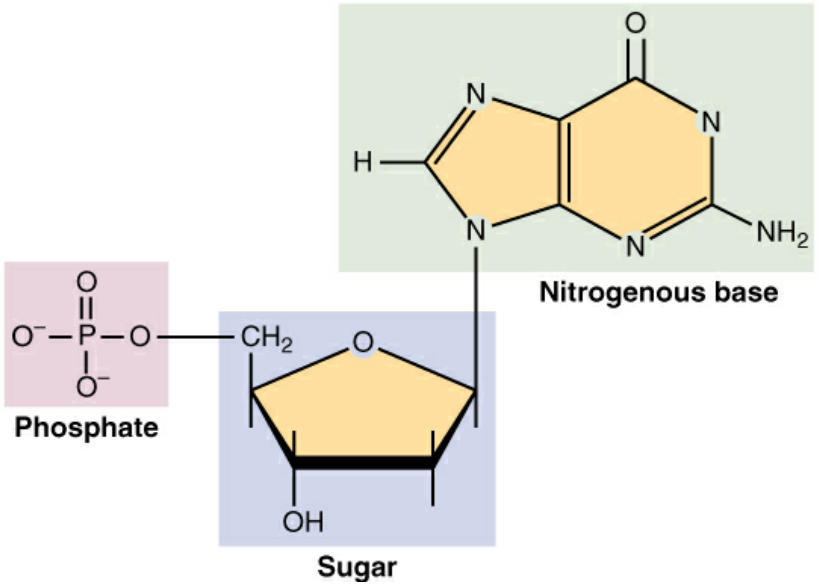


Photo 51, X-ray diffraction image of DNA

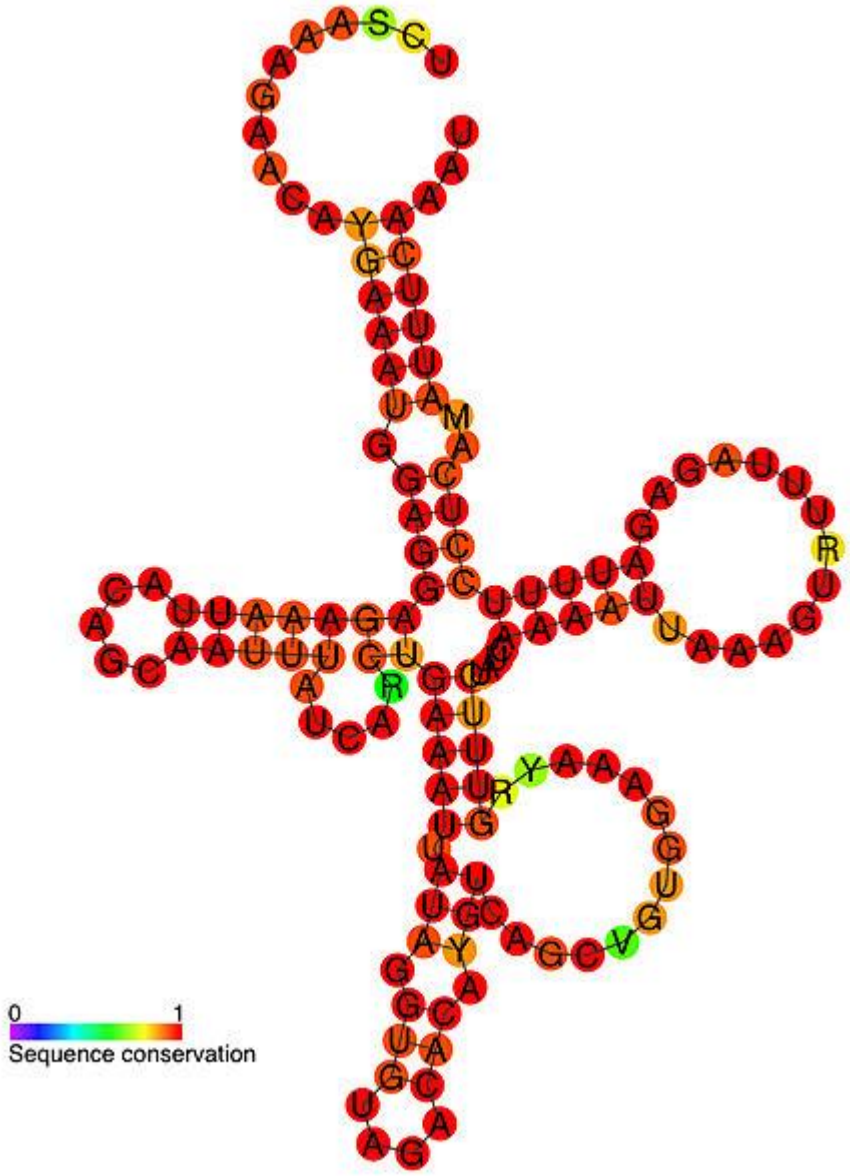
Franklin R, Gosling RG (1953) "Molecular Configuration in Sodium Thymonucleate". ***Nature*** 171**: 740–741.**





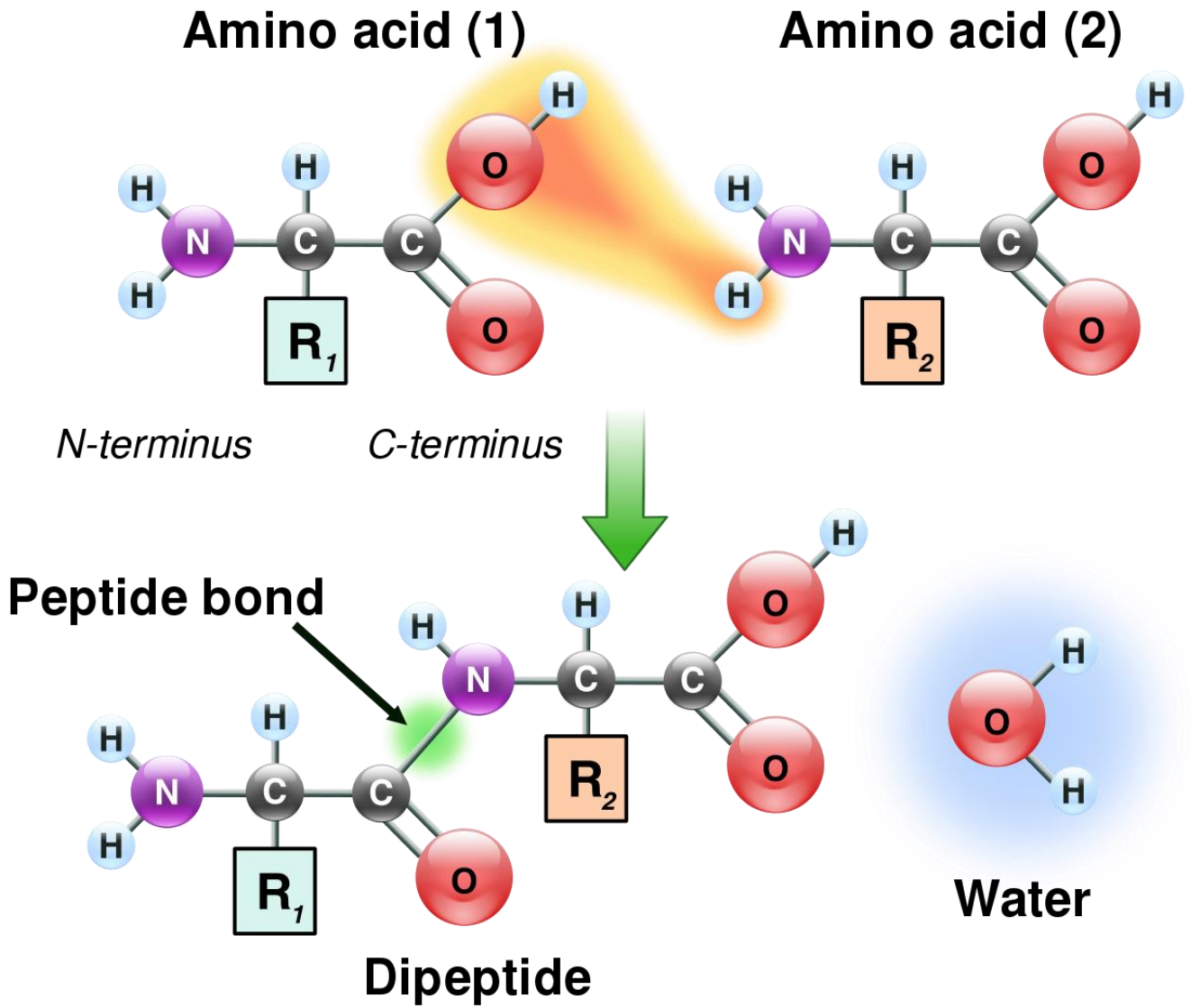
From the text book OpenStax Anatomy and Physiology, discovered through Wikimedia, reused under the CC license.

RNA structure



Downloaded from https://en.m.wikipedia.org/wiki/File%3AHAR1F_RF00635_rna_secondary_structure.jpg. Original work by wikipedia user:Ppgardne. Used under CC-SA 3.0 license.

From amino acids to proteins

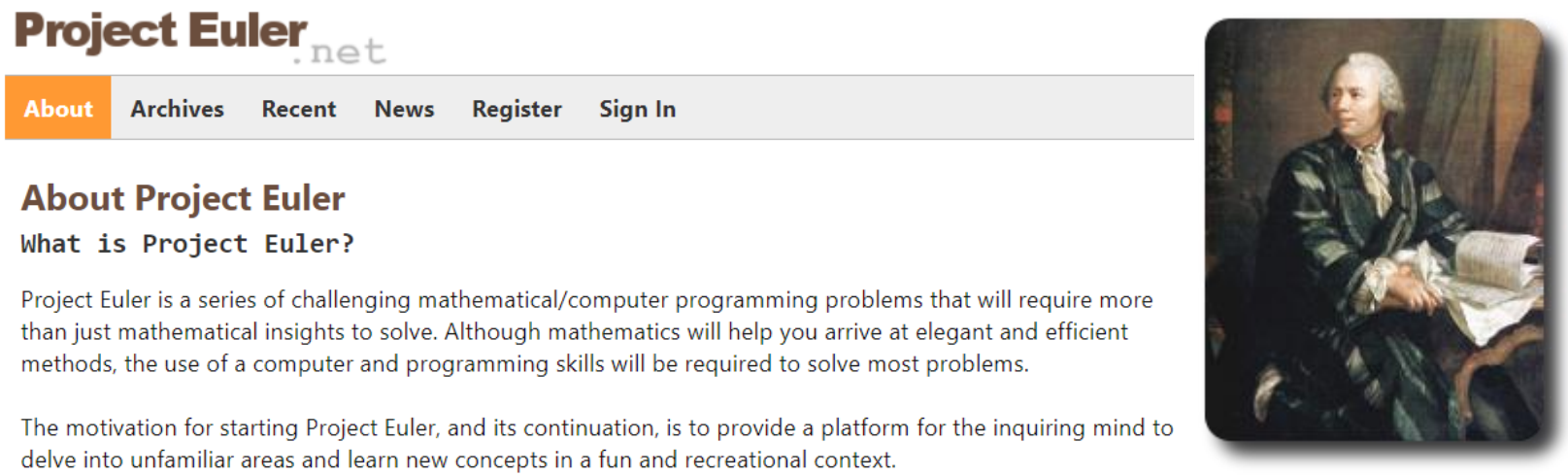


https://commons.wikimedia.org/wiki/File:Peptidformationball.svg

Software tools

- **General biological sequence analysis**
  - EMBOSS software suite: http://emboss.sourceforge.net/, also available online at European Bioinformatics Institute (EBI): https://www.ebi.ac.uk/services
  - BLAST (=Basic Local Alignment Search Tool) can be run at many places, for instances from EBI and National Center for Biotechnology Information (NCBI): https://blast.ncbi.nlm.nih.gov/Blast.cgi
  - Programming access, for instance the Biopython project: https://biopython.org
- **RNA biology**
  - ViennaRNA package (https://www.tbi.univie.ac.at/RNA/)
  - RNA processing tools available at U Bielefeld, for instance RNAhybrid, which finds minimum free energy hybridization using dynamic programming (https://bibiserv.cebitec.uni-bielefeld.de/rnahybrid)
- **Profile Hidden Markov Models (HMMs)**
  - The HMMER package: http://hmmer.org/

The Euler Project



https://projecteuler.net/

- **Learnining by problem-solving**
- **Free**
- **Math + CS**

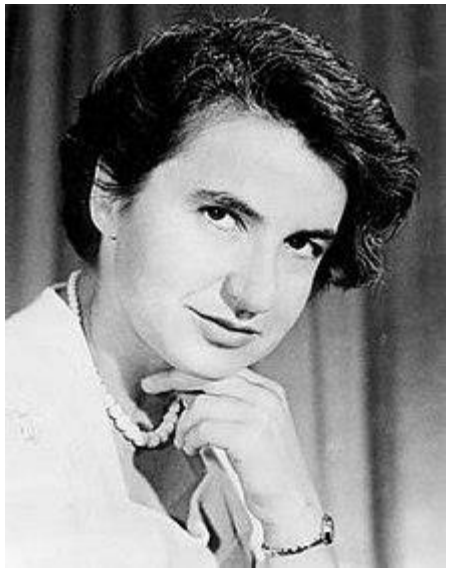
Problem 1: Multiples of 3 and 5

If we list all the natural numbers below 10 that are multiples of 3 or 5, we get 3, 5, 6 and 9. The sum of these multiples is 23.

Find the sum of all the multiples of 3 or 5 below 1000.

Rosalind: a great scientist, and a platform for learning bioinformatics and programming through problem solving

ROSALIND http://rosalind.info/problems/locations/



Rosalind Elsie Franklin

1920-1958

A Rapid Introduction to Molecular Biology click to expand

Problem

A **string** is simply an ordered collection of symbols selected from some **alphabet** and formed into a word; the **length** of a string is the number of symbols that it contains.

An example of a length 21 **DNA string** (whose alphabet contains the symbols 'A', 'C', 'G', and 'T') is "ATGCTTCAGAAAGGTCTTACG."

Given: A DNA string s of length at most 1000 nt.

Return: Four integers (separated by spaces) counting the respective number of times that the symbols 'A', 'C', 'G', and 'T' occur in s.

Sample Dataset

```
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGC
```

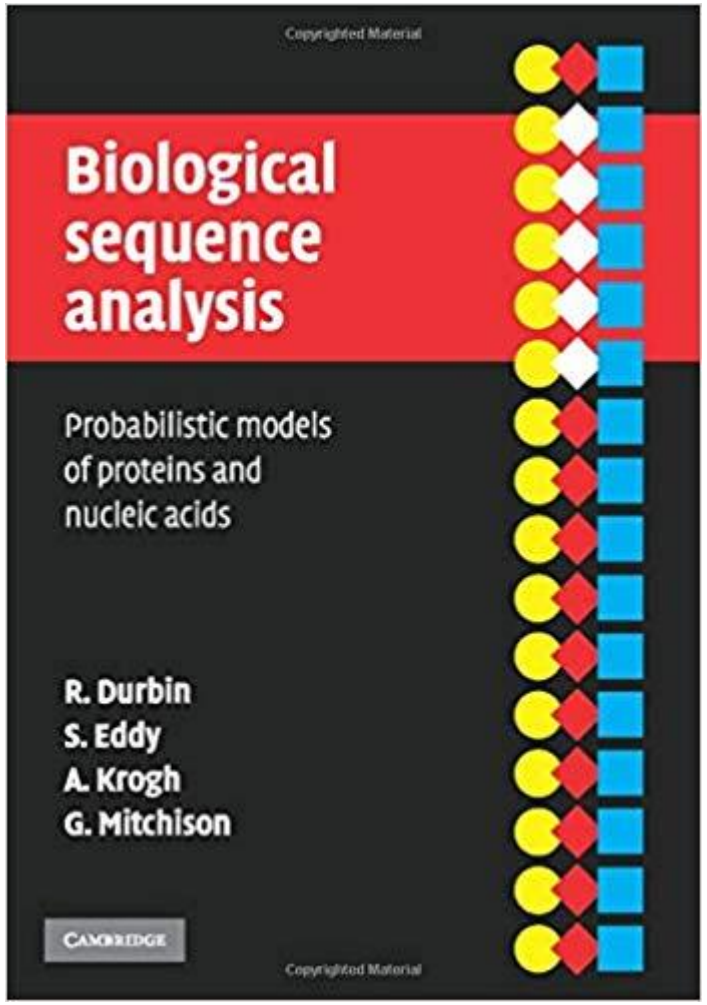
Sample Output

```
20 12 17 21
```

Please login to solve this problem.

Further resources

***Biological Sequence Analysis* by Durbin, Eddy, Krogh, and Mitchison**



[Teaching RNA algorithms](http://rna.informatik.uni-freiburg.de/Teaching/) by Backofen Lab at U Freiburg, with [source code](https://github.com/BackofenLab/RNA-Playground) available.

The website hosts among others an interative tool to visualize how dynamic programming (DP) helps to predict RNA secondary structure.

For a gentle introduction, see also *How Do RNA Folding Algorithms Work?* by Eddy, Sean R, *Nature Biotechnology* 22, Nr. 11 (November 2004): 1457–58. https://doi.org/10.1038/nbt1104-1457.

[An Introduction to Applied Bioinformatics](http://readiab.org/) by Greg Caporaso (NAU)

The tutorial is written in Python using Jupyter. It introduces concepts in (a) pairwise sequence alignment, (b) sequence homology searching, (c) generalized dynamic programming for multiple sequence alignment, (d) phylogenetic reconstruction, (e) sequence mapping and clustering, as well as (f) machine learning in bioinformatics. Applications and exercises are also available.

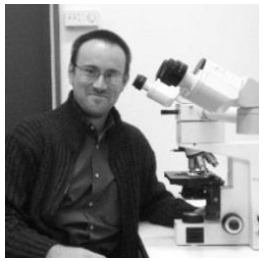
Acknowledgements





F. Hoffmann-La Roche Ltd	
Clemens Broger†	**Faye Drawnel**
Martin Ebeling	Markus Britschgi
Manfred Kansy	Roland Schmucki
Fabian Birzele	Martin Stahl
Kurt Amrein	Isabelle Wells
Annie Moisan	**Lu Gao**
Luca Piali	Lue Dai
John Young	Ravi Jagasia
Lisa Sach-Peltason	**Marco Prunotto**
Mark Burcin	**John Moffat**
Christoph Patsch	Gang Mu
Michael Reutlinger	Jianxun Jack Xie
Matthias Nettekoven	Filip Roudnicky
Andreas Dieckmann	Holger Fischer
Klas Hatje	Iakov Davydov
Laura Badi	**Ulrich Certa**
Tony Kam-Thong	Detlef Wolf
Corinne Solier	Ken Wang
Thomas Singer	Nikolaos Berntenis





External to Roche
Stefan Wiemann
Wolfgang Huber
Ozgür Sahin
Agnes Hovrat
Katharina Zweig
Sally Cowley
Alexandros Stamatakis
Michael Prummer
Mark D. Robinson
Michael Hennig
Florian Haller
Jung Kyu Canci
Verdon Taylor
Maria Anisimova
Lorenzo Gatti
Erhard van der Vries
Ab Osterhaus
Nevan Krogan
Oliv Eidam





Summary and Q&A

BACKUP

Course information

- Lecturer: Jitao David Zhang
  - jitao-david.zhang@unibas.ch (Email)
- Website: amidd.ch
- Thirteen lectures this semester
  - Introduction to drug discovery (1 session)
  - Molecular level modelling (2 sessions)
  - Omics- and cellular level modelling (2 sessions)
  - Organ- and system-level modelling (1.5 sessions)
  - Populational level modelling (1.5 sessions)
  - Case studies (1 session)
  - Invited guest speakers (2 sessions)
  - *Dies Academicus*
  - Near-end-term presentations (2 sessions)

- Fridays 12:15-14:00, two sessions of ~45 min each.
- No exercise hour yet; pre-reading and post-reading articles, as well as videos, are shared and recommended.
- We focus on interdisciplinary research with mathematics as the language and informatics as the tool.
- Both slides and board are used. Slides and notes are shared.
- The final note is given by participation (20%), presentation (30%), and an oral examination (50%).
- The oral examination will be about concepts that we learned together, and about explaining mathematical concepts (or concepts in your domain of experts) to a layman.
- **Questions**?

I am glad to share my expertise in drug discovery, and to learn from you!

Please introduce yourself!

- **Name?**
- **Background?**
- **Which part of mathematics (or other background) are you mostly interested in? Why?**
- **What do you want to take away from this course?**

Questions on the package insert info

1. What is the **indication** of *ZYRTEC*? What is its generic name?
2. What is the **gene target** of ZYRTEC?
3. How much time does ZYRTEC reaches **maximum concentration** following oral administration?
4. How long do normal vonlunteers have to **wait** until the skin wheal and flare caused by the intradermal injection of histamine is inhibited after taking 10mg ZYRTEC?
5. What types of **adverse reactions** are observed in volunteers taking ZYRTEC?
6. Is there a **biomarker** for ZYRTEC?

Questions for further thinking

- What are the commonalities between Herceptin and Zyrtec, and what are the differences?